NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

A POISSON REGRESSION ANALYSIS OF THE ACADEMIC SETBACK IN NAVAL TRAINING DEADTIME

by

Joseph T.C Li

December 2000

Thesis Advisor: Robert Read
Associate Advisor: Dennis Mar

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 2000	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE :
A POISSON REGRESSION ANALYSIS OF THE ACADEMIC SETBACK IN NAVAL TRAINING DEADTIME

5. FUNDING NUMBERS

6. AUTHOR(S)
Joseph T.C. Li

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT

The deadtime in a Naval Training Pipeline is defined as time spent by enrolled students doing things other than training. There are eight major categories of dead time and their effect has been to decrease the utilization of personnel to under 70% in recent times. Twenty-four courses for four years (1996-1999) have been selected for study. The Academic Setbacks for course with CDP identifier 6400 has been chosen for initial work and model building. The methods developed for this case will be applied to the others to the extent possible. The exploratory analyses will seek to discover internal patterns of setbacks. Failing this the process will be declared as time homogeneous and in a steady state.

14. SUBJECT TERMS
Poisson Regression, Training Deadtime, Maximum Likelihood

15. NUMBER OF PAGES
62

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

A POISSON REGRESSION ANALYSIS OF THE ACADEMIC SETBACK IN NAVAL TRAINING DEAD-TIME

Joseph T.C. Li
Lieutenant Commander, Taiwan Navy
B.S., Chinese Naval Academy, 1988

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
December 2000

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The dead time in a Naval Training Pipeline is defined as time spent by enrolled students doing things other than training. There are eight major categories of dead time and their effect has been to decrease the utilization of personnel to under 70% in recent times. Twenty-four courses for four years (1996-1999) have been selected for study. The Academic Setbacks for course with CDP identifier 6400 has been chosen for initial work and model building. The methods developed for this case will be applied to the others to the extent possible. The exploratory analyses will seek to discover internal patterns of setbacks. Failing this, the process will be declared as time homogeneous and in a steady state.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENT

The author would like to acknowledge those individuals who provided their support throughout the information-gathering phase of this thesis:

Professor Robert Read of the Operations Research Department, my thesis advisor: Thanks for your understanding, indispensable assistance and patient guidance in the whole process of the thesis.

Dennis Mar of the Naval Postgraduate School Computer Center, my second reader: For your endless hours of communicating and explaining difficult SAS statistical procedures – my most sincere thanks.

Professor Siriphong Lawphonpanich of the Operations Research Department, thanks for your unparalleled programming assistance, which helped solve the bottleneck problem and significantly contributed to my completion of this thesis.

And most importantly, to my wife Grace and my beloved son Jed, and my extended family members–Dan & Melissa Short and the saints of the Church in San Jose. Without your continuous support, encouragement, and prayers, I would not have finished this document.

I. INTRODUCTION

A. PROBLEM PRESENTATION

The deadtime in a Naval Training Pipeline refers to situations in which students are enrolled for training but not undergoing training. There are a number of reasons, e.g., waiting for a seat in a class, waiting for a transfer to the next training command or to the fleet, waiting for discharge from Naval service, or having to temporarily come out of a class once it has started. The deadtime issue has attracted more attention under the current Navy environment in which cost cutting and manpower downsizing is emphasized, especially since its effect has decreased the utilization of personnel to under 70% in recent times. Reducing deadtime is beneficial for both the Navy, which pays for it, and the sailor who endures it

B. RESEARCH BACKGROUND AND PRIMARY QUESTION

Two studies are cited for background. Belcher (1999) considers student not-under instruction time and reveals the impact and contribution of eight major categories of dead time. Belcher's document analyzes causes and recommends methods to decrease the time awaiting instruction (AI), awaiting training (AT) and instruction interruption (II). In another study, Rhoades (1998) suggests information systems for integrating the Navy's recruiting, training, and assignment in order to optimize the entire system.

These above studies identified deadtime and its cost to the Navy. Another important issue is that of identifying those time periods during a course of instruction that

experience the beginning of deadtime. This thesis develops a method to identify these deadtime bottlenecks.

C. THESIS OUTLINE

The next chapter, Methodology, introduces the foundation of Poisson regression. The third chapter describes the given data. Chapter IV, Model Fitting, fits the Poisson regression model to the selected data and calculates the deviance as measure of goodness of fit. The following Chapter V computes, interprets, and analyzes the output, and reveals the usefulness of the output as well. Chapter VI concludes the work with some recommendations.

D. EXPECTED BENEFITS OF THIS THESIS

Using a Poisson regression analysis, we intend to locate the worst deadtime bottleneck in a particular course. To simplify the analysis, we consider the Academic Setbacks of course 6400 for model building and exploratory data analysis. The analysis will reveal any deadtime bottlenecks that should be identified and considered for possible administrative action. If there are no bottlenecks, we will declare the process as time homogeneous and in a steady state, requiring no adjustment. The methodology used in this study can be extended to other Navy courses to identify significant deadtime categories.

II. METHODOLOGY

A. INTRODUCTION

Poisson regression analysis is appropriate for response variables that have non-negative integer values: 0, 1, 2.... The Poisson distribution is used to describe the response; the behavior of the mean value function in various categories is the goal of modeling.

The occurrences of deadtime events of the type in our study are relatively rare. Let's examine the Academic Setback of the course 6400. The number of student academic setbacks must be 0, 1, 2... the non-negative integer values. One student academic setback is assumed to be independent of any other student academic setback. The total number of academic setbacks for a single course are not large, but there are many courses, and the overall problem becomes large.

B. POISSON DISTRIBUTION

The Poisson distribution has a single parameter; called lambda, λ, which is the average or expected number of events per unit of time, i.e. the mean μ. Interestingly, the variance of the Poisson distribution is also equal to λ. The values possibly taken by the Poisson random variable are the non negative integers.

The mathematical expression of the Poisson distribution for obtaining y events, given that λ events are expected, is

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \qquad 2.1$$

where Y = the Poisson random variable.

P = the probability of y events given a knowledge of λ.

λ = expected number of counts, i.e., the mean μ.

e = the base of the natural logarithm (approximated by 2.71828).

y = user supplied input.

C. THE METHOD OF MAXIMUM LIKELIHOOD

The method of maximum likelihood for the estimation of statistical parameters is the one used in this thesis. This method selects the value of λ, based upon the data, which maximizes the likelihood function of the observed results. A likelihood function takes positive values. Often it is easier to work with the log-likelihood function than the likelihood function itself. Since the logarithmic function is a monotonically increasing function, the estimator that maximizes the log-likelihood function will maximize the likelihood function as well. This log likelihood function takes negative values.

D. THE POISSON REGRESSION PROCEDURE

The Poisson regression procedure hypothesizes a model to explain the observed data. The maximum likelihood method is used to estimate the parameters of the model.

The most general case of a Poisson regression (the saturated model) defines an individual λ for each data point in a sample of size N:

$$L(y;\lambda)=\prod_{i=1}^{N}\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) \qquad 2.2$$

A Poisson regression model would define some relationship among the λ_i:

$$L(y;\lambda^*)=\prod_{i=1}^{N}\left(\frac{e^{-\lambda_i}\lambda_i^{*y_i}}{y_i!}\right) \qquad 2.3$$

and λ_i is estimated by y_i in the saturated model.

E. THE MEASURES OF GOODNESS OF FIT

One measure of how closely the Poisson regression model fits the observed data is called the deviance $D(\lambda_e^*)$ for the regression model (Kleinbaum, page 503):

$$D(\lambda_e^*)=-2\ln\left[\frac{L(y;\lambda_e^*)}{L(y;\lambda)}\right] \qquad 2.4$$

where $L(y;\lambda_e^*)$ is the estimated likelihood of the proposed model,

and $L(y;\lambda)$ is that of the saturated model.

The better the Poisson regression model fits the observed data, the closer $L(y;\lambda_e^*)/L(y;\lambda)$ gets to one. Since both the numerator and denominator are maximum likelihood estimators, the $D(\lambda_e^*)$ statistic is approximately a chi-square variate with N-K degrees of freedom. K is the number of parameters in the model.

THIS PAGE INTENTIONALLY LEFT BLANK

III. STATISTICAL DESCRIPTION OF DATA

The given data include 24 courses from 1996 to 1999. Our first task is to explore the selected data sets for their statistical properties. Table 3.1 reveals the structure of the data for a single course, a single category of dead time raw data, which consist of eight

Table 3.1. The Abstract of Raw Data of Academic Setback from 1996 to 1999.

OBS	ENROLL FY	CIN	CDP	CRSEL EN	CATEGOR Y	ABBRNM	COUNT	DAY
1	1996	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK READ SKL	1	114
2	1996	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK READ SKL	1	119
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	1996	A-041-0010	6400	164	AS	STBK ACAD WITHT RTRNG ADMIN	1	98
49	1997	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK CMPRHN SUBJ MATR	1	120
50	1997	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK CMPRHN SUBJ MATR	1	14
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
125	1997	A-041-0010	6400	164	AS	STBK ACAD WITHT RTRNG ADMIN	1	97
126	1998	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK CMPRHN SUBJ MATR	1	100
127	1998	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK CMPRHN SUBJ MATR	1	65
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
152	1998	A-041-0010	6400	164	AS	STBK ACAD WITHT RTRNG ADMIN	1	96
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
157	1999	A-041-0010	6400	164	AS	STBK ACAD RTRNG CLSRM LACK CMPRHN SUBJ MATR	1	97

columns.

These eight columns contain: the fiscal year (ENROLLFY), the course identification notation (CIN), the category of deadtime (CATEGORY), the course number (CDP), the course length (CRSELEN), the day number into the course (DAY) of the event, the numbers of students entering deadtime on that day (COUNTS), and the deadtime reason (ABBRNM). For simplification of the initial work and model building, we selected the Academic Setback in CDP 6400. Therefore the CATEGORY, the CDP and the CRSELEN are AS, 6400 and 164 days and are fixed in this example. We focus on the DAY and the COUNTS of the events.

Table 3.2 organizes the information into how many events happened for each day into the course. For instance, four students received academic setbacks on the 14^{th} day of the course, one student on the 99^{th} day, but no students on the 19th day, etc. If there was no student setback on a day, such as the 19^{th}, the original data set did not include a record for DAY = 19. Thus the day 19 does not appear. We did find some students having a setback after 164 days. These are viewed as miss-entries and are ignored.

Table 3.3 records the frequency of the various COUNTS. It records the number of days for each category of COUNTS. For example, out of 164 days, there were no setbacks declared on 70 of the days and there were exactly one on 35 of the days, etc. The variable COUNTS in Table 3.3 takes on eight values: zero, one, two, three, four, five, six and seven. The total proportion of counts for one, two and three is 49.38%, while the total for four, five, six and seven is 7.93%.

Table 3.2. Frequency Table by DAY.

DAYS	Frequency (COUNTS)	Percent	DAYS	Frequency (COUNTS)	Percent	DAYS	Frequency (COUNTS)	Percent
14	4	1.80	59	4	1.80	99	1	0.45
15	4	1.80	61	6	2.70	100	3	1.35
16	3	1.35	62	1	0.45	101	2	0.45
17	1	0.45	63	3	1.35	102	1	0.45
18	1	0.45	64	1	0.45	103	2	0.90
20	7	3.15	65	2	0.90	104	1	0.45
21	1	0.45	67	2	0.90	105	2	0.90
22	2	0.90	68	1	0.45	111	3	1.35
23	1	0.45	69	4	1.80	113	1	0.45
25	3	1.35	70	1	0.45	114	1	0.45
28	1	0.45	71	2	0.90	115	1	0.45
31	3	1.35	72	2	0.90	118	2	0.90
32	1	0.45	74	2	0.90	119	1	0.45
33	2	0.90	75	1	0.45	120	4	1.80
34	1	0.45	77	2	0.90	121	3	1.35
35	4	1.80	78	3	1.35	123	3	1.35
36	2	0.90	80	4	1.80	124	3	1.35
37	1	0.45	82	2	0.90	125	3	1.35
38	6	2.70	83	3	1.35	133	5	2.25
39	7	3.15	84	2	0.90	137	1	0.45
40	5	2.25	85	3	1.35	138	1	0.45
42	3	1.35	86	2	0.90	139	1	0.45
45	2	0.90	87	3	1.35	143	2	0.90
46	2	0.90	88	3	1.35	146	1	0.45
47	1	0.45	89	3	1.35	148	1	0.45
48	2	0.90	90	1	1.35	159	1	0.45
49	3	1.35	91	2	0.90	167	1	0.45
52	1	0.45	92	3	1.35	171	1	0.45
53	3	1.35	93	2	0.90	172	2	0.90
54	1	0.45	94	1	0.45	173	1	0.45
55	3	1.35	95	2	0.90	174	1	0.45
56	2	0.90	96	1	0.45	181	2	0.90
57	1	0.45	97	2	1.35	183	1	0.45
58	1	0.45	98	1	0.45	220	1	0.45

Table 3.3. Frequency Table by COUNTS.

COUNTS	Frequency	Percent
0	70	42.68
1	35	21.34
2	25	15.24
3	21	12.80
4	7	4.27
5	2	1.22
6	2	1.22
7	2	1.22
Sum	164	100.00

Figure 3.1[1] graphically illustrates the distribution of the setbacks. To obtain the general trend in the data, we group seven day sets into weeks (ten days in the last period). Since there was not enough data for 1999, it was pooled with 1998 data for display.

The peaks suggest time bottlenecks marking the student setbacks. In both of the years 1996 and 1997, the peak value happened in the sixth week, but in the year 1998+99 it did not. Rather than having a common peak location for each year, the peaks move back and forth. At this point, we do not know whether the cyclical effect is real, or merely an artifact of randomness. The graphs may be misleading, because these peaks move when we change the size of the grouping.

Figure 3.1 The Distribution of AS-6400 in COUNTS per Week

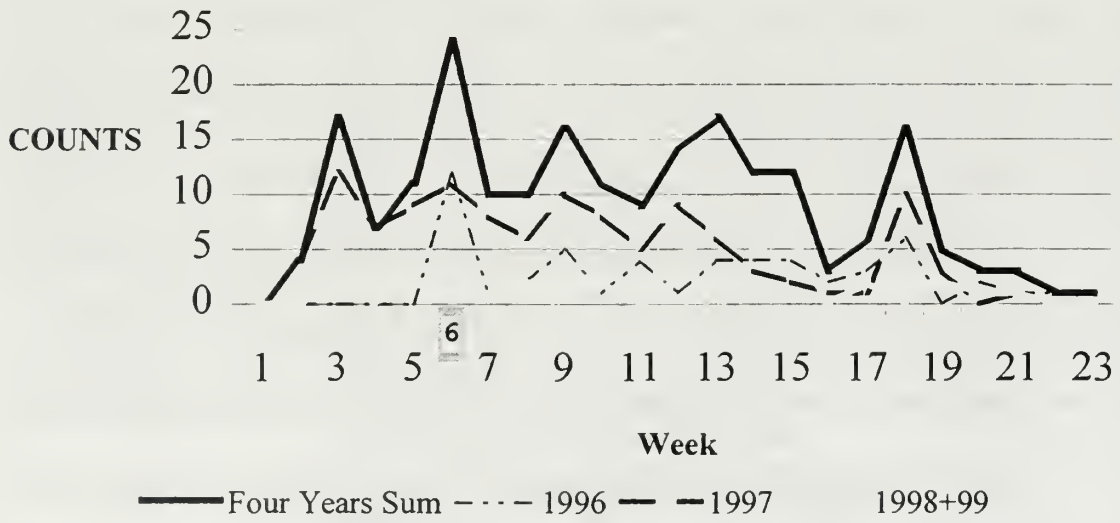

Figure 3.1. The Distribution of AS-6400 in COUNTS per Week from 1996-to 1999.

[1] The data in year 1999 included five observations only, therefore we combine them into the year 1998 and mark as 1998+99.

IV. MODEL FITTING

A. MODEL DESCRIPTION

We describe the observed setbacks using DAY as the sole explanatory variable. The sequence of course days is divided into K intervals. Within an interval, the λ is the same for each day's Poisson distribution. Between intervals, the λs can be different.

For example, if K=3 and suppose the total number days in a course is N=150, the first interval could contain the first 60 days. The second interval could include the next 40 days, and the last piece includes the remaining 50 days. A different partition might use intervals of lengths 34, 57, and 59.

The parameter λ is the expected value of the response variable in a Poisson process. It is the rate of the counts on a day and can be repersented as $\lambda = \lambda(\text{days})$ by a mean value function, that is the function of the explanatory variables. Since the λs are the same for each day in an interval, the maximum likelihood estimator of the interval's λ (the Academic Setback rate) is equal to the sum of the setbacks divided by the number of days in the interval.

To illustrate the idea of the explanatory variables DAY and interval, consider Table 4.1 below, Figures 3.1, 4.1 and the output of program **read.gam**, which we will discuss later. In Table 4.1, we sum the counts to get a response from the DAY 14 to 125, which is around the 3rd to the 18th week in the Figure 3.1 as well as the 2nd interval in Figure 4.1. Then we divide the responses by the number of days in the interval

Table 4.1. The COUNTS Corresponding to the DAY, Interval and Week.

DAY	Int	Wk	COUNT	DAY	Int	Wk	COUNT	DAY	Int	Wk	COUNT	DAY	Int	Wk	COUNT
1			0	43			0	85			3	127			0
2			0	44			0	86			2	128			0
3			0	45			2	87			3	129	3		0
4		1	0	46		7	2	88		13	3	130		19	0
5			0	47			1	89			3	131			0
6			0	48			2	90			1	132			0
7	1		0	49			3	91			2	133	4		5
8			0	50			0	92			3	134			0
9			0	51			0	93			2	135			0
10			0	52			1	94			1	136			0
11		2	0	53		8	3	95		14	2	137		20	1
12			0	54			1	96			1	138			1
13			0	55			3	97			2	139			1
14			4	56			2	98			1	140			0
15			4	57			1	99			1	141			0
16			3	58			1	100			3	142			0
17			1	59			4	101			2	143			2
18		3	1	60		9	0	102		15	1	144		21	0
19			0	61	2		6	103	2		2	145			0
20			7	62			1	104			1	146			1
21			1	63			3	105			2	147			0
22			2	64			1	106			0	148	5		1
23			1	65			2	107			0	149			0
24			0	66			0	108			0	150			0
25		4	3	67		10	2	109		16	0	151		22	0
26			0	68			1	110			0	152			0
27			0	69			4	111			3	153			0
28	2		1	70			1	112			0	154			0
29			0	71			2	113			1	155			0
30			0	72			2	114			1	156			0
31			3	73			0	115			1	157			0
32		5	1	74		11	2	116		17	0	158			0
33			2	75			1	117			0	159		23+	1
34			1	76			0	118			2	160			0
35			4	77			2	119			1	161			0
36			2	78			3	120			4	162			0
37			1	79			0	121			3	163			0
38			6	80			4	122			0	164			0
39		6	7	81		12	0	123		18	3				
40			5	82			2	124			3				
41			0	83			3	125			3				
42			3	84			2	126	3		0				

Figure 4.1 The Five-Interval Policy of AS-6400 in 1996--1999

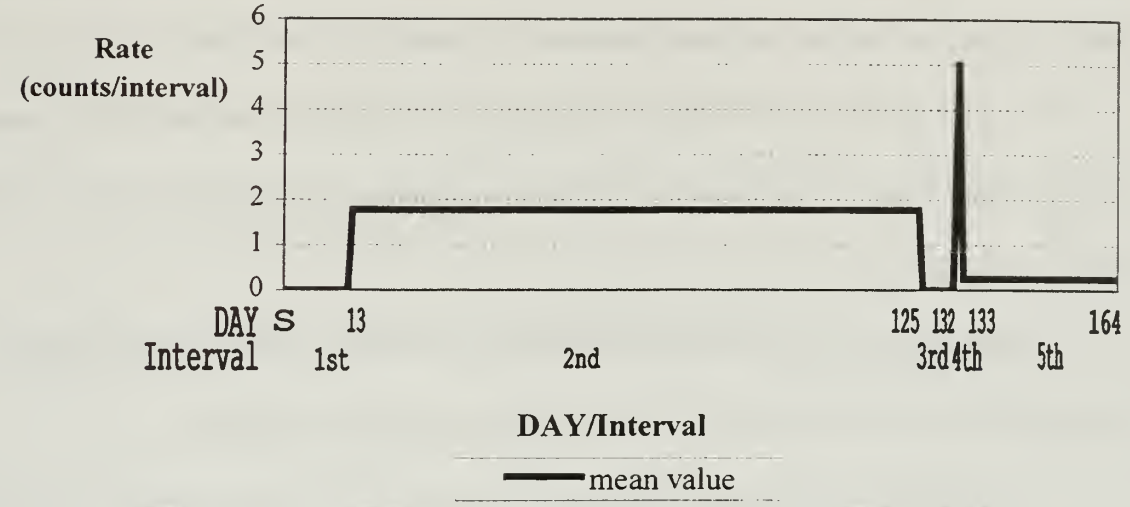

Figure 4.1. The Five-Interval Policy of AS-6400 in 1996-1999.

to get the average rate.

Note that the rate in Figure 3.1 varies from week to week but in Figure 4.1 it is a constant within the chosen intervals. A week, being a fixed interval that slices the course length mechanically, disguises the trend of the curve. Consequently, we model λ(days) as a simple step function and choose breakpoints by maximum likelihood. This will follow the trend with variable length intervals, so that the λs are constant over well-selected intervals of days.

B. EXAMPLE

The course with CDP number 6400 is used as a tangible example to illustrate the method. The course has N = 164 days. Let Y be a Poisson random response for a particular day. Thus Y is the number of setbacks (the variable COUNTS) observed that

day. The N days are partitioned into K intervals. It is convenient to describe the partition by a set of breakpoints b_j, j=1...K. The breakpoints are the indices of the last day in each interval where the days are numbered consecutively from day 1 to day N.

The observed values of Y from Table 3.1 are $y_1 = 0$, $y_{16} = 3$, and such. The course is N=164 days long. Let K=5 for a five interval partition. Pick breakpoints at 13, 125, 132, 133, and 164.

Under these conditions we can calculate the number of days in each partition: p_1=13, p_2=125-13=112, p_3=132-125=7, p_4=133-132=1, and p_5=164-133=31.

To calculate the maximum likelihood estimator λ_j for an interval, sum the observed number of setbacks and divide by the number of days in the interval. For example, for interval 5, the sum of occurrences is 1+1+1+2+1+1+1=8. The maximum likelihood estimator of λ_5 is 8/31=0.258. Note that for both intervals 1 and 3, the λ estimate is 0.

C. METHOD

Let $Y_1...Y_N$ be N independent Poisson random variables, one for each day in the course. The days in the course are partitioned into K intervals and all of the Poisson variables associated within an interval have a common parameter.

The partition can be described in two ways. Both are given because some of the equations are greatly simplified by using one notation or the other.

Take $p_1...p_k$ where p is the number of days in an interval. The sum of the p_1 through p_k is N.

We take $b_1 \ldots b_k$ as the breakpoints of each interval. The b_j is the index of the last element in interval j. The index starts from day 1. Note that $b_1 = p_1$, $b_2 = p_1 + p_2$, and so on. The last breakpoint, b_k, equals to N. To simplify notation later on, we define $b_0 = 0$.

The likelihood function of the saturated model is

$$L(y;\lambda) = \prod_{i=1}^{N} P(y;\lambda) = \prod_{i=1}^{N} \left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right) \qquad 4.1$$

The log-likelihood function of the saturated model is

$$\ln L(y;\lambda) = -\sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N} y_i \ln \lambda_i + \sum_{i=1}^{N} y_i! \qquad 4.2$$

For the Poisson regression model we have K intervals. Within an interval the Poisson distribution for each day uses the same λ. Let λ^* be the K sets of λs of the regression model, then the likelihood function of the regression model is

$$L(y;\lambda^*) = \prod_{j=1}^{K} \prod_{i=b_{j-1}+1}^{b_j} \frac{e^{-\lambda_j} \lambda_j^{y_i}}{y_i!} \qquad 4.3$$

The log-likelihood of the regression model has the complex form

$$\ln L(y;\lambda^*) = -\sum_{j=1}^{K} p_j \lambda_j + \sum_{j=1}^{K} \left[(\ln \lambda_j) \left(\sum_{i=b_{j-1}+1}^{b_j} y_i \right) \right] + \sum_{i}^{N} y_i! \qquad 4.4$$

We want to select the partition, $p_1 \ldots p_k$, which maximizes the log-likelihood of the regression model. Identifying best partition is computer intensive and is accomplished using a program implementing the network shortest path algorithm. This was accomplished with the help of Dennis Mar of the Systems Management Department and

Professor Lawphonpanich of the Operations Research Department. An outline of the method is presented in Appendixes A, B, and C.

The measures of goodness of fit of Poisson regression models are obtained from the comparison of maximized likelihood values. We use the deviance to produce likelihood ratio tests for assessing the goodness of fit.

The log-likelihood ratio statistic has the form:

$$D(\lambda^*) = -2\ln\left[\frac{L(y;\lambda^*)}{L(y;\lambda)}\right] \qquad 4.5$$

where $D(\lambda^*)$ is the deviance for the regression model

If the model is a valid one, then the $D(\lambda^*)$ statistic has approximately a chi-square distribution with N-K degrees of freedom. As the log-likelihood of the regression model increases, the deviance statistic decreases. If the deviance is large and the chi-square distribution test rejects the null hypothesis that the step function model is tenable. This is evidence that the regression model does not fit.

V. COMPUTATION AND ANALYSIS

A. COMPUTATION

This chapter traces the methodology of fitting the models to the data, and judging goodness of fit. The steps are to maximize the log likelihood (or minimize the negative of the log likelihood as that quantity is more directly related to the deviance statistic), and use the Chi-square distribution to judge goodness of fit. Three programs are utilized: two programs **read2.gam** and **read.gam** were created by Professor Lawphonpanich and a SAS program was developed by Dennis Mar. The guidance for using these programs is explained in the Appendix E.

The **read2.gam** (Appendix A) calculates the negative maximum log likelihood for any K value specific interval policy, e.g. a three intervals policy, five intervals...etc. Its output provides us output about the start numbers of intervals for that policy.

The **read.gam** (Appendix B) finds the best choice of contiguous intervals, such as (13, 112, 17, 1, 31) for a five interval policy of AS in course 6400 from fiscal year 1996 to 1999. These are the interval lengths; the break points are DAY 13, 125, 142, 143, and 173. This is the best choice for a five interval policy.

The SAS program includes procedures **Data Step** and **Deviance & Alpha (α) Value Step**. The data step selects and organizes data from the raw data and prepares formatted data to the **gam** programs. Then the deviance & alpha value step computes the

deviance and calculates the chi-square test statistic distribution upper tail confidence level.

Tables 5.1, 5.2, 5.3, and 5.4, consolidating the outputs of AS course 6400 in 1998 from those three programs, are used to interpret the output, illustrate goodness of fit, and analyze the output.

B. INTERPRETATION OF OUTPUT

Table 5.1 is the output of the **read2.gam**. The values in the first column signify the number of intervals in the policy: three to 10. The values in the second column, is the negative of the maximum log likelihood value. In the way that **read2.gam** calculates, smaller is better. Comparing the difference of the value between three and four (2.59), four and five (8.64), five and six (2.74) interval policy, the pair four and five has the biggest change. This biggest marginal value in these three pairs suggests the five interval policy is plausibly a reasonable first choice for the next step.

Table 5.1 The Negative Maximum Log Likelihood of Various Interval Policy.
for AS6400_98

	Value
3	84.8867
4	82.2978
5	73.6616
6	70.9237
7	67.9204
8	66.5098
9	62.5889
10	59.9181

Table 5.2 includes output of program **Read.gam** and SAS **Deviance & Alpha Value Step**.

The **Read.gam** output shows the negative of the maximum log likelihood, breakpoints, indicators and average rate of the partition. The negative of the maximum log likelihood value is 73.6616, which is the same as Table 5.1, the five interval policy. The first column noted as [s .D14], [D14 .D16] is the beginning and the ending day of interval. S is the start day of the course and the D14 is the 14th day of the course...etc. The 1st interval is from s to D14, the 2nd, is from D14 to D16...etc. The breakpoints are $(b_1, b_2, b_3, b_4, b_5) = (14, 16, 77, 105, 164)$ and the lengths of intervals are $(p_1, p_2, p_3, p_4, p_5) = (14, 2, 61, 28, 59)$.

Table 5.2 The Best Five Intervals, Deviance and α Value of Chi-Square Test.

```
Five Partitions Policy
VARIABLE  TOTCOST.L        =    73.6616  negative log likelihood
PARAMETER output
                      X            A
s    .D14           1.0000         ...
D14  .D16           1.0000       2.5000
D16  .D77           1.0000       0.1475
D77  .D105          1.0000       0.7857
D105 .D164          1.0000

The SAS System Output of 6400_AS_98 for 5 Partition Deviance & Alpha Value
Obs     dev        Alpha
1     155.831    0.55624
```

The X column is a binary variable, which indicates whether an interval was selected for the final model: 1 is for selected and 0 is not. The zeros do not appear in the output. In our formulation of the problem only the included intervals are showed. The X value is always 1 therefore we can ignore it.

The A column is the average rate (λ) for the interval. A missing value in the A column implies 0. The average Academic Setback rate of partition [D14 .D15] is 2.5 counts for each day during this period of time.

In the SAS output of deviance and alpha value, the deviance is 155.831 and the α value is 0.55624 with 159 degrees of freedom (=164 – 5, course length minus numbers of intervals). The α value will be discussed more in the Goodness of Fit section.

C. GOODNESS OF FIT

Under the null hypothesis, H_o: K partition model fits the observed data. The distribution of the deviance statistic is Chi-square. Let α be the probability that the deviance random variable is greater than or equal to the realized deviance statistic. At the 5% level of significance, calculated values of α greater than 0.05 supports the null hypothesis. Consider five, three and seven interval policy first (Tables 5.2, 5.3 and 5.4).

The alpha (α) value indicates the probability that we would observe a deviance value of that size or smaller when the null hypothesis is true. This alpha level is 0.06381

Table 5.3 The Best Three Intervals, Deviance and α Value of Chi-Square Test.

```
AS6400_98_1
Three Partitions Policy
VARIABLE  TOTCOST.L       =    84.8867  negative log likelihood
PARAMETER output

                  X           A
s   .D77        1.0000      0.1818
D77 .D105       1.0000      0.7857
D105.D164       1.0000      0.1017

The SAS System output
6400_AS_98 for 3 Partition Deviance & Alpha Value
Obs    dev        Alpha
1    187.005    0.06381
```

for the best three interval policy, 0.55624 for the best five-interval policy, and 0.82357 for the best seven interval policy.

Table 5.4 The Best Seven Intervals, Deviance and α Value of Chi-Square Test.

```
Seven Partition Policy
VARIABLE  TOTCOST.L       =      67.9204  negative log likelihood
PARAMETER output
                    X           A
s    .D14        1.0000
D14  .D16        1.0000       2.5000
D16  .D32        1.0000
D32  .D33        1.0000       2.0000
D33  .D77        1.0000       0.1591
D77  .D105       1.0000       0.7857
D105.D164        1.0000       0.1017

The SAS System Output
6400_AS_98 for 7 Partition Deviance & Alpha Value
Obs      dev        Alpha
 1     140.482    0.82357
```

The results lead to a basic dilemma. How many intervals are suitable for the analysis and requisite recommendations? This is a trade-off between the number of intervals K and the goodness-of-fit statistic. The three-interval policy is desirable because of its simplicity. But while its deviance value would not be rejected at the .05 levels, it is close. The practitioner could reasonable select between the five- and seven-interval policies. For the remainder of this thesis, the seven-interval policy will be studied.

D. ANALYSIS OF OUTPUT

We construct Figure 5.1 from the output of the seven interval policy for analysis due to its higher confidence level. Referring to the figure, the 2nd, the 4th and the 6th

intervals attract our attention more than others. This seven interval policy follows a low-high pattern. The rates of setback are 0.00, 2.50, 0.00, 2.00, 0.16, 0.79, and 0.10. The 1st and the 3rd intervals consisting of 14 days and 16 days were near the beginning of the course,where low setback rates are expected. The 2nd and the 4th consisting of two and one days may reflect the learning problems from previous intervals which were not dealt with until those particular days. Looking beyond the first four intervals, the last three show up in the three interval style that we prefer.

Figure 5.1 The Seven-Interval Policy for AS of Course 6400 in 1998+99

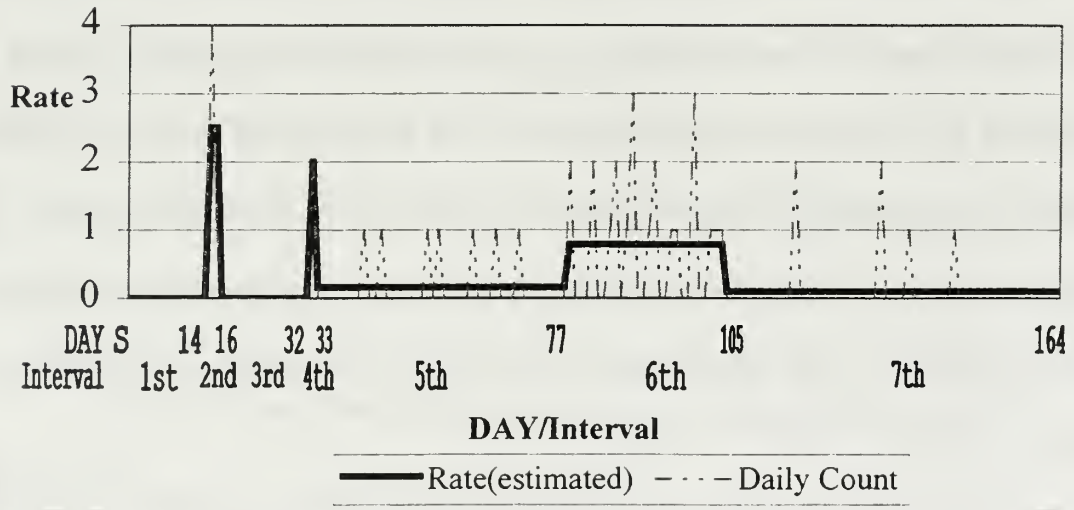

Figure 5.1. The seven-interval policy of AS-6400 in 1998+.

Following the low-high pattern, the last three intervals exhibit a specified style: increasing interval – high setback rate interval – decreasing interval. The 5th interval included 44 days, the 6th interval 28 days and the 7th interval 59 days. The 6th interval

E. ADVANTAGES OF OUTPUT

Compare the daily count to the interval rate on the figure 5.1, the interval rate simplifies the curve and our study. Taking advantage, we try to use the seven interval policy to be our general policy. Its use is to identify common periods of time having a commonality of concerns in the course.

Figure 5.2 consolidates the seven interval policy over the four years. The four-year sum curve accumulates all contributions from the four years of data and displays some stable rates except for the 3rd interval (day 39 to 40) and the 6th interval (day 133). The rates in the 1st interval (0.00), the 5th (0.00) and the 7th (0.25) are approximately equal as well as the rate in the 2nd (1.75) and the 4th (1.6353).

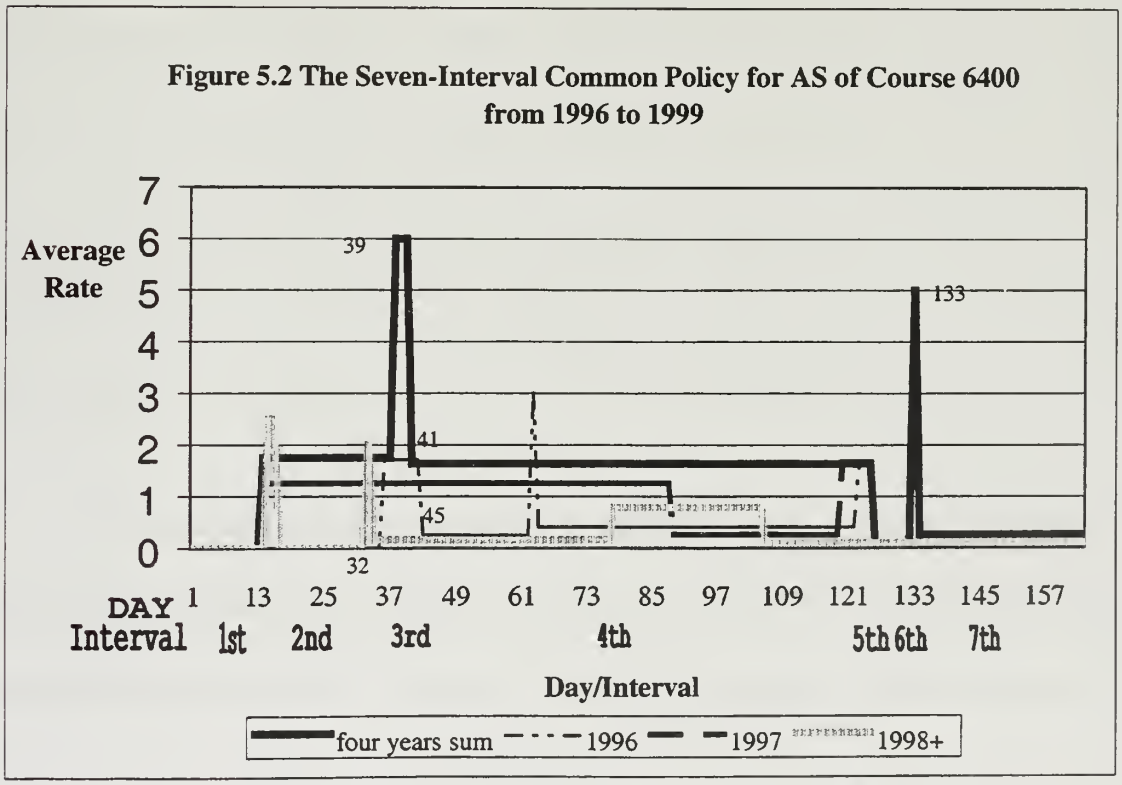

Figure 5.2. The Consolidated Common Policy of AS-6400.

The rates in the 1st interval (0.00), the 5th (0.00) and the 7th (0.25) are approximately equal as well as the rate in the 2nd (1.75) and the 4th (1.6353).

The consolidated figure gives us an advantage to know the common behavior of the intervals when we compare them from year to year. Obviously the 3rd interval in the four-year sum curve catches more significance than the 5th interval. At nearly the same period of time, from day 32 to day 45, the rate rises promptly. This stage includes the 2nd interval for year 1997 with a stable rate 1.2581 in the long term day 13-88, the 2nd interval for year 1996 with a rate 1.7143 during day 35-42 and the 4th interval with a rate 2 on the day 32 for year 1998+99. Therefore, the days 32 to 45 of the course will have the first priority for administrative attention.

VI. CONCLUSIONS

To illustrate further how the developed method works; we test it on Academic Attrition and Instruction Interruption (exclude Holiday Leave) of the course 6400. The Table 6.1 displays the result.

Table 6.1. The Test Output of AA and II in Course 6400.

```
Read.gam output for AA6400_mix
                                        82.0508 negative log likelihood
                    X           A
s  .D21         1.0000
D21.D91         1.0000      0.1857
D91.D164        1.0000      0.2877


The SAS System output for AA_6400_m
Obs    dev         Alpha
1    144.834     0.78277

The SAS System output for course II6400_m 7 partition
Obs    dev         Alpha
1    536.684          0

The SAS System output for course II6400_96 7 partition
Obs    dev         Alpha
1    588.022          0

The SAS System output for course II6400_97 7 partition
Obs    dev         Alpha
1    591.873          0

The SAS System output for course II6400_98 7 partition
Obs    dev         Alpha
1    471.576          0

The SAS System output for course II6400_99 7 partition
Obs    dev         Alpha
1    363.443          0
```

The three interval policy for academic attrition achieves $\alpha = 78\%$. We do not to reject the null hypothesis for Academic Attrition. The 3rd interval is the interesting period with the highest attrition rate. Turning to Instruction Interruption, the test rejects ($\alpha = 0$)

for the seven interval policy and all years. One must use more intervals for Instruction Interruption in course 6400.

A. CONCLUSIONS

This study has applied the developed method on three categories of deadtime and in course 6400 from 1996 to 1999: the Academic Setback, the Academic Attrition and the Instruction Interrupt; and concludes with three results. First, the bottlenecks happen approximately around the 32^{nd} to 45^{th} day for the academic setback. Second, the bottlenecks happen on 92^{nd} and 93^{rd} days for the academic attrition with 78% confidence. Third, the Instruction Interrupt needs more than seven intervals to reach a satisfactory Chi-square test. It is a candidate for the time homogeneous process.

The developed estimators detect the location of the weaknesses by course and category of deadtime for a given data set. Since so many courses are taught and so many categories of deadtime exist, it's not possible to locate all of the possible problems with a common model. Finding the location of the weakest point is always the first priority.

Recall that we assumed the deadtime incidence rate is constant if the course is in a stable status. The developed estimator calculates rates from best choice intervals. Long intervals indicate stability, short ones suggest a transient nature. This type of instability needs further research.

B. RECOMMENDATIONS

This thesis is a pilot study and the developed estimator provides a flexible tool for the task. The possible future studies include:

- Develop a user-friendly program which executes the same function as this thesis.
- Analyze the contribution and relationship of the reasons to deadtime in the concerned interval.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIXES

APPENDIX A: READ2.GAM

This program calculates Maximum Log Likelihood value. The bold head prints should be changed according to the length of the course, the name of the data file and the numbers of interval for policies.

```
$TITLE *Gameside's Program for policy Log Maximum Likelihood
*-------------------------------------------------------------
$OFFUPPER OFFSYMLIST OFFSYMXREF INLINECOM{ }
OPTIONS RESLIM = 900, ITERLIM = 100000
        LIMCOL = 0, LIMROW = 0, DECIMALS = 4, SOLPRINT = OFF
        OPTCR = 0.05
        LP = OSL; {**OSL has a network solver**}
*-------------------------------------------------------------
Set
   i           /s, D1*D164/


Parameters
   y(i)    /
$include as6400_D_98.prn
/;
*-------------------------------------------------------------
Set arc(i,i);

Alias (i,j,k);

Scalar NGrp ;

Parameters
   a(i,j)  optimal a
   c(i,j)  obj value
   b(i)    ;

a(i,j)$(ord(j) gt ord(i)) =
          sum(k$(ord(k) gt ord(i) and ord(k) le
ord(j)),Y(k))/(ord(j) - ord(i));


c(i,j)$(ord(j) gt ord(i)) = (ord(j) - ord(i))*a(i,j)
```

```
             - sum(k$(ord(k) gt ord(i) and ord(k) le
ord(j)),Y(k))*log(a(i,j))$
                                (a(i,j) gt 0);
arc(i,j) = YES$(ord(j) gt ord(i));

b(i) = 1$(ord(i) eq 1) -1$(ord(i) eq card(i));

*display c;

*POSITIVE VARIABLE
Binary Variable
  X(i,j)  amount of flows on each arc;

VARIABLE
  TOTCOST  negative log likelihood;
*-------------------------------------------------------------
EQUATIONS
   OBJ         define objective function
   FLOWBAL(i)  flow conservation
   numint;
*-------------------------------------------------------------
OBJ..
      TOTCOST =E= SUM((i,j)$arc(i,j),c(i,j)*X(i,j));
FLOWBAL(I)..
    SUM(j$arc(i,j),X(i,j))-SUM(j$arc(j,i),X(j,i))=E=b(i);
NUMINT..
    sum((i,j)$arc(i,j), X(i,j)) =L= ngrp;
*-------------------------------------------------------------
MODEL MCFLOW / ALL /;

Set iter /1*365/;

Parameter report(*,*), sol(i,j);
Scalar old /99999/;

Loop(iter$(ord(iter) ge 3 and ord(iter) le 10),
     ngrp = ord(iter);
     solve mcflow using MIP minimizing TOTCOST;
     report(iter,'Value') = Totcost.l;
     IF (totcost.l lt (old-0.0001),
         sol(i,j) = X.L(i,j);
         old = totcost.l;
        );
    );
option Y:2:0:1;
display report;
```

APPENDIX B: READ.GAM

This program selects the best breakpoints of the interval under the maximum log likelihood value. The bold head prints should be changed according to the length of the course, the name of data file, and the number of interval.

```
$TITLE *  * Gamside's Program for Five Pieceses Policy * * *
*-------------------------------------------------------------
$OFFUPPER OFFSYMLIST OFFSYMXREF INLINECOM{ }
OPTIONS RESLIM = 900, ITERLIM = 100000
        LIMCOL = 0, LIMROW = 0, DECIMALS = 4, SOLPRINT = OFF
        OPTCR = 0.05
        LP = OSL; {**OSL has a network solver**}
*-------------------------------------------------------------
Set
   i           /s, D1*D164/


Parameters
   y(i)    /
$include as6400_D_98.prn
/;
*-------------------------------------------------------------
Set arc(i,i);

Alias (i,j,k);

Scalar NGrp ;

Parameters
   a(i,j)  optimal a
   c(i,j)  obj value
   b(i)    ;

a(i,j)$(ord(j) gt ord(i)) =
          sum(k$(ord(k) gt ord(i) and ord(k) le
ord(j)),Y(k))/(ord(j) - ord(i));


c(i,j)$(ord(j) gt ord(i)) = (ord(j) - ord(i))*a(i,j)
             - sum(k$(ord(k) gt ord(i) and ord(k) le
ord(j)),Y(k))*log(a(i,j))$
                                  (a(i,j) gt 0);
arc(i,j) = YES$(ord(j) gt ord(i));
```

```
b(i) = 1$(ord(i) eq 1) -1$(ord(i) eq card(i));

*display c;

*POSITIVE VARIABLE
Binary Variable
  X(i,j)  amount of flows on each arc;

VARIABLE
  TOTCOST  negative log likelihood;
*-------------------------------------------------------------
EQUATIONS
   OBJ         define objective function
   FLOWBAL(i)  flow conservation
   numint;
*-------------------------------------------------------------
OBJ..
      TOTCOST =E= SUM((i,j)$arc(i,j),c(i,j)*X(i,j));
FLOWBAL(I)..
    SUM(j$arc(i,j),X(i,j))-SUM(j$arc(j,i),X(j,i))=E=b(i);
NUMINT..
    sum((i,j)$arc(i,j), X(i,j)) =L= ngrp;
*-------------------------------------------------------------
MODEL MCFLOW / ALL /;

ngrp = 5;
SOLVE MCFLOW USING MIP MINIMIZING TOTCOST;
DISPLAY TOTCOST.L;

parameter output(i,j,*);

output(i,j,'X') = X.L(i,j);
output(i,j,'A')$(X.L(i,j) = 1) =
      sum(k$(ord(k) gt ord(i) and ord(k) le
ord(j)),Y(k))/(ord(j) - ord(i));

display output;
```

APPENDIX C: SAS PROGRAM

1. Data Step

The **Data Step** of this program selects the desired data from the raw data. The bold head prints should be changed according to the deadtime CATEGORY, CDP and ENROLLFY of desired data, the course length, degree of freedom, and the breakpoints.

```
Data Step
*********************;
**** Select data ****;
*********************;
data data1;
  set diskh.spbsnum;
  if category='AS';
  if cdp     ='6400';
  if enrollfy=1998;
*****************************************;
**** Count number of setbacks each day ****;
*****************************************;
Proc freq data=data1 noprint;
  table days / out=data2;
  weight count;
*********************************************************;
**** Remove any data for any day greater than the ****;
**** maximum length of the course.                ****;
*********************************************************;
data data2;
  set data2;
  if days>164 then delete;
  drop percent;
*********************************************************;
**** Create a data set where each observation is a day***;
**** of the course.                                   ***;
*********************************************************;
data data3;
  do days=1 to 164;
    output;
  end;
*********************************************************;
**** Add in the days with count=0 setbacks.          ****;
**** "data3" has an entry for each day of the course. ****;
*********************************************************;
```

```
data diskh.alldays;
  merge data2 data3;
  by days;
  if count=. then count=0;
*************************************************************;
*Create a list of log factorials,ln(0!) through ln(100!),*;
*************************************************************;
proc print;
run;
```

2. Deviance & Alpha Value Step

The Deviance & Alpha Value Step of this program has to run with the Data Step together to calculate the deviance and look up the Alpha value of Chi-square test statistical distribution. The bold head prints should be changed according to the course length, degree of freedom, and the breakpoints.

```
Deviance & Alpha Value Step

************************************************************
****  The incoming data set contains 164 observations.  ***
****  The variable COUNT for the ith observation is the ***
****  total number of setbacks for the ith training day.***
************************************************************
**  The transpose procedure changes the arrangement of the*
**  data set.  The 164 observations of COUNT are converted*
**  into a new data set with one observation              *
**  and 164 variables d1 through d164.                    *
**  The value, for example, of d34 is equal to the value  *
**  of COUNT in the 34th observation.                     *
************************************************************
****  This reconfiguration is done soley because of the   *
****  style of syntax used by SAS.                        *
************************************************************;
proc transpose data=diskh.alldays out=transp prefix=d;
  var count;
****
****  The variables are added: n (total days in course),
****  df (degrees of freedom of the chi-square,
```

```
****  p1 p2 p3 p4 p5 (number of days in each piece of the
partition).
****;
data allpart;
  set transp;
  drop _name_;
  n  = 164;
  df = 159;
  p1 = b1;                b1 = 14;
  p2 = b2-b1;             b2 = 16;
  p3 = b3-b2;             b3 = 77;
  p4 = b4-b3;             b4 = 105;
  p5 = b5-b4;             b5 = 164;
*************************************************************;
**  The ultimate goal in this data step is calculation of *;
**  the "deviation" for the partition specified by p1     *;
**  through p5.                                           *;
*************************************************************;
data allpart;
  set allpart;
  array ff(i) f1-f164;
  array mm(i) m1-m164;
*************************************************;
******  Calculate average setbacks of piece 1.*;
*************************************************;
  sumx=0;
  do i = 1 to p1;
    sumx=sumx+ff;
  end;
  meanx=sumx/p1;
  do i = 1 to p1;
    mm=meanx;
  end;
  ****;*****************************************;
  ****  Calculate average setbacks of piece 2. *;
  *************************************************;
  sumx=0;
  do i = p1+1 to p1+p2;
    sumx=sumx+ff;
  end;
  meanx=sumx/p2;
  do i = p1+1 to p1+p2;
    mm=meanx;
  end;
  *************************************************;
  ****  Calculate average setbacks of piece 3. *;
  *************************************************;
```

```
sumx=0;
do i = p1+p2+1 to p1+p2+p3;
  sumx=sumx+ff;
end;
meanx=sumx/p3;
do i = p1+p2+1 to p1+p2+p3;
  mm=meanx;
end;
************************************************;
****  Calculate average setbacks of piece 4.  *;
************************************************;
sumx=0;
do i = p1+p2+p3+1 to p1+p2+p3+p4;
  sumx=sumx+ff;
end;
meanx=sumx/p4;
do i = p1+p2+p3+1 to p1+p2+p3+p4;
  mm=meanx;
end;
************************************************;
****  Calculate average setbacks of piece 5. *;
************************************************;
sumx=0;
do i = p1+p2+p3+p4+1 to n;
  sumx=sumx+ff;
end;
meanx=sumx/p5;
do i = p1+p2+p3+p4+1 to n;
  mm=meanx;
end;
*******************************************;
**** Calculate deviation which is -2 times *;
**** the log of the ratio of the likelihood*;
**** of the hypothesized model and the     *;
**** likelihood of the saturated model     *;
*******************************************;
dev=0;
do i = 1 to 164;
  if ff LT 1.e-15 then dev= dev - (ff-mm);
                  else dev= dev + ff*log(ff/mm);
end;
dev = 2*dev;
*****************************************************;
****  Calculate the cumulative probability for the *;
****  chi-square distribution from 0 to dev        *;
****  for df degrees of freedom.                   *;
*****************************************************;
```

```
  alpha=1-probchi(dev,df);
  drop m1-m164 f1-ff164;
*************************************************;
****  Print the deviance and the Alpha.        *;
*************************************************;
proc print data=allpart;
  var dev alpha;
title "Five piece partition, category=AS cdp=6400
       enrollfy=1998";
run;
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D: TEST RESULTS

```
AS6400_m_7P
VARIABLE  TOTCOST.L            =      90.6703  negative log likelihood
PARAMETER output
                    X           A
s    .D13        1.0000
D13  .D37        1.0000      1.7500
D37  .D40        1.0000      6.0000
D40  .D125       1.0000      1.6353
D125.D132        1.0000
D132.D133        1.0000      5.0000
D133.D164        1.0000      0.2581


The SAS System Output of AS_6400_mix for 7 Partition Dev & Alpha value
Obs      dev      Alpha
 1    341.205        0


AS6400_96_7P
VARIABLE  TOTCOST.L            =      80.4718  negative log likelihood
PARAMETER output
                    X           A
s    .D35        1.0000
D35  .D42        1.0000      1.7143
D42  .D62        1.0000      0.2500
D62  .D63        1.0000      3.0000
D63  .D122       1.0000      0.4068
D122.D125        1.0000      1.6667
D125.D164        1.0000      0.1282


The SAS System Output of AS6400_96_7P for 7 Partition Dev & Alpha value
Obs      dev      Alpha
 1    155.153   0.5267


AS6400_97_7P
VARIABLE  TOTCOST.L            =     100.6404  negative log likelihood
PARAMETER output
                    X           A
s    .D13        1.0000
D13  .D88        1.0000      1.2533
D88  .D119       1.0000      0.2581
D119.D125        1.0000      1.6667
D125.D132        1.0000
D132.D133        1.0000      3.0000
D133.D164        1.0000      0.0323


The SAS System Output of AS6400_97_7P for 7 Partition Dev & Alpha value
Obs      dev      Alpha
 1    249.702        0
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E: GUIDANCE for USING PROGRAMS

1. Preparing Data:

a. Selecting and organizing desired data (**AS6400_98**, Academic Setback of course 6400 in 1998) by SAS Program **Data Step** (Appendix C) to produce the following two dimensions output.

The SAS System Data Step Output from AS6400_98

DAY	COUNT	DAY	COUNT	DAY	COUNT	DAY	COUNT	DAY	COUNT
1	0	34	0	67	0	100	3	133	2
2	0	35	0	68	0	101	0	134	0
3	0	36	0	69	1	102	0	135	0
4	0	37	0	70	0	103	1	136	0
5	0	38	0	71	0	104	1	137	0
6	0	39	0	72	0	105	1	138	1
7	0	40	0	73	0	106	0	139	0
8	0	41	0	74	0	107	0	140	0
9	0	42	1	75	0	108	0	141	0
10	0	43	0	76	0	109	0	142	0
11	0	44	0	77	0	110	0	143	0
12	0	45	1	78	2	111	0	144	0
13	0	46	0	79	0	112	0	145	0
14	0	47	0	80	0	113	0	146	1
15	4	48	0	81	0	114	0	147	0
16	1	49	0	82	2	115	0	148	0
17	0	50	0	83	0	116	0	149	0
18	0	51	0	84	0	117	0	150	0
19	0	52	0	85	1	118	2	151	0
20	0	53	1	86	2	119	0	152	0
21	0	54	0	87	0	120	0	153	0
22	0	55	1	88	1	121	0	154	0
23	0	56	0	89	3	122	0	155	0
24	0	57	0	90	0	123	0	156	0
25	0	58	0	91	0	124	0	157	0
26	0	59	0	92	1	125	0	158	0
27	0	60	0	93	2	126	0	159	0
28	0	61	1	94	0	127	0	160	0
29	0	62	0	95	0	128	0	161	0
30	0	63	0	96	1	129	0	162	0
31	0	64	0	97	1	130	0	163	0
32	0	65	1	98	0	131	0	164	0
33	2	66	0	99	0	132	0		

b. Reorganizing the previous SAS Data output and saving in Formatted Text (Space delimited) format for programs **read2.gam** and **read.gam**.

The input data of **AS6400_D_98.prn** for both read2.gam and read.gam

D1	0	D34	0	D67	0	D100	3	D133	2
D2	0	D35	0	D68	0	D101	0	D134	0
D3	0	D36	0	D69	1	D102	0	D135	0
D4	0	D37	0	D70	0	D103	1	D136	0
D5	0	D38	0	D71	0	D104	1	D137	0
D6	0	D39	0	D72	0	D105	1	D138	1
D7	0	D40	0	D73	0	D106	0	D139	0
D8	0	D41	0	D74	0	D107	0	D140	0
D9	0	D42	1	D75	0	D108	0	D141	0
D10	0	D43	0	D76	0	D109	0	D142	0
D11	0	D44	0	D77	0	D110	0	D143	0
D12	0	D45	1	D78	2	D111	0	D144	0
D13	0	D46	0	D79	0	D112	0	D145	0
D14	0	D47	0	D80	0	D113	0	D146	1
D15	4	D48	0	D81	0	D114	0	D147	0
D16	1	D49	0	D82	2	D115	0	D148	0
D17	0	D50	0	D83	0	D116	0	D149	0
D18	0	D51	0	D84	0	D117	0	D150	0
D19	0	D52	0	D85	1	D118	2	D151	0
D20	0	D53	1	D86	2	D119	0	D152	0
D21	0	D54	0	D87	0	D120	0	D153	0
D22	0	D55	1	D88	1	D121	0	D154	0
D23	0	D56	0	D89	3	D122	0	D155	0
D24	0	D57	0	D90	0	D123	0	D156	0
D25	0	D58	0	D91	0	D124	0	D157	0
D26	0	D59	0	D92	1	D125	0	D158	0
D27	0	D60	0	D93	2	D126	0	D159	0
D28	0	D61	1	D94	0	D127	0	D160	0
D29	0	D62	0	D95	0	D128	0	D161	0
D30	0	D63	0	D96	1	D129	0	D162	0
D31	0	D64	0	D97	1	D130	0	D163	0
D32	0	D65	1	D98	0	D131	0	D164	0
D33	2	D66	0	D99	0	D132	0		

2. Calculating Maximum Negative Log Likelihood

The program **read2.gam** (Appendix A) uses the formatted data **AS6400_D_98.prn** as input to calculate the maximum negative likelihood of different policy. The following is an example output.

The output of **read2.gam** for **as6400_D_98.prn**

	Value
3	84.8867
4	82.2978
5	73.6616
6	70.9237
7	67.9204

```
8       66.5098
9       62.5889
10      59.9181
```

3. Selecting Best Choice of Partition

The program **read.gam** (Appendix B) uses the formatted data **AS6400_D_98.prn** as input to select best policy combination. The following is an example output.

```
The output of read.gam for as6400_D_98.prn

VARIABLE  TOTCOST.L       =    73.6616  negative log likelihood
PARAMETER output
                    X           A
s    .D14        1.0000
D14  .D16        1.0000      2.5000
D16  .D77        1.0000      0.1475
D77  .D105       1.0000      0.7857
D105 .D164       1.0000
```

4. Calculating Deviance & Alpha Value

The SAS program (Appendix C) **Deviance & Alpha Value Step** use first column of **read.gam** output which is the index of best choice as input to calculate deviance and look up Alpha value. The following is an example output.

```
The SAS System Output of AS6400_98
for 5 Partition Deviance & Alpha Value

Obs     dev       Alpha
1    155.831    0.55624
```

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Allison, Paul, *LOGISTIC REGRESSION USING THE SAS SYSTEM Theory and Application*, SAS Institute Inc., 1999

Belcher, Steven W., *Analysis of Student Not-Under-Instruction Time in Initial Skills Training: Trends, Causes, and Proposed Fixes*, Center for Naval Analyses, Jan 1999

Delwiche, D. L. and Slaughter, J.S., *The Little SAS Book,Second edition*, Cary, NC SAS Institute Inc.,1998

Dobson, Annette J., *An Introduction to Generalized Linear Models*, T. J. Press (Padstow) Ltd, 1990

Kleinbaum, David G., *Applied Regression Analysis and Other Multivariable Methods* PWS-KENT Publishing Company, 1988

Levine, M. D., Berenson, L. M. and Stephan, D., *Statistics for Managers using Microsoft Excel, Second Edition*

Render, B. and Stair M. R. JR., *Quantitative Analysis for Management, Seventh Edition* Pewnrixw-Hall Inc, 2000

Rhoads, Cathleen Marie, *Dead Time In Naval Training*, Master's Thesis, University of Maryland, Sep 1998

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center 2
   8725 John J. Kingman Road, Ste 0944
   Fort Belvoir, VA 22060-6218

2. Dudley Knox Library 2
   Naval Postgraduate School
   411 Dyer Road
   Monterey, California 93943-5101

3. Professor Rober Read, Code OR/Re 1
   Naval Postgraduate School
   1 University Circle
   Monterey, CA 93943-5001

4. Professor Siriphong Lawphongpanich, Code OR/Lp 1
   Naval Postgraduate School
   1 University Circle
   Monterey, CA 93943-5001

5. Dennis Mar, Code SM/Mn 1
   Naval Postgraduate School
   555 Dyer Road
   Monterey, CA 93943-5101

6. Porfessor Roger Evered, Code SM/Ev 1
   Naval Postgraduate School
   555 Dyer Road
   Monterey, CA 93943-5101

7. LTCDR Li, Tien-Cheng 6
   15 Lane 142 Chung-Ching S. Rd. Sec. 1
   Taipei, Taiwan 106